MODELING OF GRBAS PERCEPTUAL EVALUATION USING SPECTRAL FEATURES OBTAINED FROM AN AUDITORY-BASED FILTERBANK

R. Fraile[1], K. Neumann [2], J.M. Gutiérrez-Arriola[1], N. Sáenz-Lechón[1], V.J. Osma-Ruiz[1]

[1]Signal Theory & Communications Department, Universidad Politécnica de Madrid, Madrid, Spain
[2]Department of Phoniatrics & Paedaudiology, St. Elisabeth Hospital, Ruhr-University Bochum, Bochum, Germany
rfraile@ics.upm.es, Katrin.Neumann@ruhr-uni-bochum.de, jmga@ics.upm.es,
nslechon@ics.upm.es, vosma@ics.upm.es

Abstract: **Perceptual voice evaluation according to the GRBAS scale is modelled using a linear combination of acoustic parameters calculated after a filter-bank analysis of the recorded voice signals. Modelling results indicate that for breathiness and asthenia more than 55% of the variance of perceptual rates can be explained by such a model, with only 4 latent variables. Moreover, the greatest part of the explained variance can be attributed to only one or two latent variables similarly weighted by all 5 listeners involved in the experiment. Correlation factors between actual rates and model predictions around 0.6 are obtained.**

Keywords: **Perceptual evaluation, Linear modelling, Auditory models.**

I. INTRODUCTION

Since the primary function of human voice is interpersonal communication, voicing is closely related to hearing. For this reason, many protocols for voice quality assessment currently in use include perceptual evaluation of voice quality [1]. The widespread use of standardized scales such as GRBAS [1] or CAPE-V [2] has contributed to increasing the value of perceptual rating as a clinical tool, in spite of the reliability issues identified by some researchers [3].

Specifically, GRBAS has been recommended as a minimum standard for perceptual evaluation in the voice clinic [4]. It includes the evaluation of five aspects of voice: overall grade (G), roughness (R), breathiness (B), asthenia (A), and strain (S). For each one, the rater has to assign a mark ranging from 0 (best quality) to 3 (worst). In general, G tends to be easier to evaluate that R, B, A or S [5], "easier" meaning that a lower degree of variability is to be expected, both inter-rater and intra-rater.

Regarding reliability of GRBAS, values around 0.6 have been reported for inter-rater Pearson's correlation coefficient in scales G, R and B [1]. Reliability may also be measured in terms of the Cohen's kappa statistic [5] but, since inter-rater agreement is greater when continuous scales are used [6], correlation coefficients are to be preferred for evaluating such agreement in authors' view.

Due to the limited reliability of perceptual evaluation of voice, the search for objective acoustic descriptors of voice quality has received a great deal of attention in the scientific community for years. The relation between these descriptors and perceptual ratings of voice have also been investigated. For instance, correlations of G, R, B, and A with noise parameters have been found [7]. Similarly, correlations of R and B with both noise measures and pitch/amplitude perturbation measures have been reported [8]. Some spectral measures have also been proposed and they have shown to provide relevant correlations mainly with B [9]. A low-dimensional coding of the overall spectral shape in cepstral domain has shown to provide fair correlations with R and B [10] and the cepstral peak prominence (CPP) also exhibits significant correlations with G, R, and B [11].

This paper presents an analysis of the perceptual rates assigned by five raters to the voice of 47 individuals according to the GRBAS scheme. Inter-rater correlations of rates corresponding to the same scale are studied. The relationship between this set of rates and voice measures obtained after a filter-bank analysis similar to that presented in [12] is also studied. Such processing scheme models the front end of the auditory system and, consequently, it is expected to provide acoustic measures that are relevant to perception. The conclusions in [13] prevent against the use of measures describing the spectral shape and other researchers also point out that the temporal dynamics of the outputs of filters in the filter-bank are more relevant to perception than the overall spectral shape, even when calculated in short-time frames [14]. Consequently, the acoustic measures used here intend to provide simple, low-level descriptions of such dynamics.

The obtained results indicate that for B and A up to 55% of the variance of the perceptual rates can be explained by a few factors combining these low-level measures, and that most of such variance can be explained by factors that are common and similarly weighted for all raters.

II. MATERIALS

Voice recordings corresponding to 20 patients (14 females, 6 males) and 27 healthy speakers (15 females, 12 males) were available. Average age for female patients

was 45.3 while for female healthy speakers it was 40.5. Similarly, for male patients the average age was 57.2 and for healthy speakers it was 36.9.

All voices were recorded in the Phoniatrics & Paedaudiology Department of the St. Elisabeth Hospital (Ruhr-University Bochum) in a quiet room within a normal clinical environment. No special attempt was made to prevent the appearance of background noises. Recordings were collected at a sampling rate equal to 22,050 Hz and with 32 quantization bits using a system from XION-Medical (XION GmbH). All recordings were normalised to have a root mean square value equal to 1.

All patients and healthy speakers were asked to pronounce at comfortable pitch and intensity. A head-mounted microphone was used in order to keep distance between lips and microphone constant (approximately 20 cm). The recording used in this investigation corresponds to the reading of a German translation of Aesop's fable "The northwind and the sun".

III. Methods

A. Perceptual evaluation

Recordings corresponding to the reading of Aesop's fable were presented to five listeners with diverse levels of experience in voice evaluation. Specifically, the listeners were a phoniatrician with 17 years of experience (the author KN) and four advanced bachelor students of logopedics (in their 6th semester). All of them were asked to assign labels according to the GRBAS scales to each recording. Label values for all five scales were allowed to vary between 0 and 3 with a resolution equal to 0.25, though the students kept resolution of their labels coarser (0.5). All five listeners held a joint meeting for training and discussion before performing the GRBAS evaluation.

B. Processing of voice recordings

From the recordings, the first sentence of the fable was selected: "*Einst stritten sich Nordwind und Sonne, wer von ihnen beiden wohl der Stärkere wäre, als ein Wanderer, der in einen warmen Mantel gehüllt war, des Weges kam*". As a pre-processing stage, intervals corresponding to voiced sounds were selected, according to the algorithm described in [15] with prior μ-law compression so as to attenuate the highest peaks.

Recording segments corresponding to voiced sounds were processed by a filter-bank consisting of 22 filters with pass bands corresponding to the 22 first auditory critical bands, as detailed in [12]. However, instead of Hamming-based filters, gammatone responses were preferred since they provide a better model for the front end of the auditory system. Specifically, the implementation proposed by Slaney was used [16].

The dynamics of the filter-bank output signals were described in terms of two parameters defined in [12]: the average energy for each band and the band energy decorrelation time, which is a measure of the stability of the signal energy (longer decorrelation times correspond to more stable signal energies). Consequently, for each recording 44 parameters were obtained, corresponding to the average energy and the energy decorrelation time at the output of each one of the 22 filters.

C. Statistical analysis

Correlations between GRBAS rates and between rates and acoustic parameters were measured in terms of the Spearman correlation coefficient. This measure was preferred instead of the more common Pearson correlation coefficient because of its capability for measuring non-linear relations between variables. The correction for ties proposed in [17, chap. 5] was implemented in the computation of the correlation coefficients.

Modelling of the GRBAS rates as linear combinations of the values of the acoustic parameters was done by means of partial least squares (PLS) regression [18]. If $\mathbf{Y}$ is a 47×5 matrix containing the rates assigned by each one of the five listeners to each recording and corresponding to either G, R, B, A or S, and $\mathbf{X}$ is a $47\times N_p$ matrix containing the values of a selected set of N_p out of the available 44 acoustic parameters associated to each voice recording, then the PLS model approximates the rates as:

$$\mathbf{Y} \approx \mathbf{X} \cdot \mathbf{F} \cdot \mathbf{W} \qquad (1)$$

where $\mathbf{F}$ is a $N_p \times N_f$ matrix that reduces the dimensionality of the space of input variables from N_p down to N_f. N_f is the number of latent variables or factors of the model. $\mathbf{W}$ is a $N_f \times 5$ matrix that models the weight that each listener assigns to each one of the factors in order to generate the corresponding rates.

Inputs and outputs of the PLS model were linearised and normalised following standard procedures [18]. Namely, acoustic parameters were transformed according to a fourth-root law (a substitute for the logarithmic law when null or almost null values occur) and mean subtraction and variance normalisation were applied to both transformed acoustic parameters and GRBAS rates.

IV. Results

A. Inter-rater correlations

Tab. I shows the values of the Spearman coefficients measuring correlation between rates corresponding to the same scale and assigned by different listeners.

As expected, the values of the correlation coefficients for scale G are larger than for the rest of scales. This is consistent with results from other researchers [5]. The values for the Spearman correlation coefficients are also greater than the Pearson coefficients reported in [1], but this is also as expected because Pearson coefficients are not sensitive to non-linear relations.

Table I. Minimum, maximum and average inter-rater correlation coefficients for all five scales in GRBAS.

Scale	Minimum correlation	Maximum correlation	Average
G	0.73	0.88	0.80
R	0.66	0.80	0.75
B	0.55	0.84	0.73
A	0.65	0.87	0.78
S	0.39	0.84	0.64

Table II. Positive and negative relevant correlations between GRBAS rates and acoustic parameters. *AE* stands for *average energy* and *DT* stands for *decorrelation time*. f_c is the central frequency of the corresponding critical band in Hz.

f_c	G		R		B		A		S	
	AE	*DT*	*AE*	*DT*	*AE*	*DT*	*AE*	*DT*	*AE*	*DT*
60								+		
150										
250				–						
350										
455	–		–		–		–			
570		–			–	–	–			–
700		–		–						
845										
1000					–					
1175	–	–	–	–	–	–	–	–	–	–
1375	–				–		–		–	
1600					–		–			
1860					–	–	–	–		
2160		–		–	–	–	–	–		–
2510										
2925										
3425	–									
4050										
4850					+		+		+	
5850		–		–	+					
7050					+					
8600					+		+		+	

B. Identification of the relevant acoustic parameters

In order to identify which acoustic parameters among the 44 available ones were the most relevant for modelling GRBAS rating, the following procedure was implemented. Firstly, the correlation coefficients between each parameter and the rates corresponding to each scale and each listener were calculated. This resulted in a set of 44×5×5=1100 values. Secondly, the 90-percentile of the absolute values of these correlation coefficients was set as a threshold for selection. Last, all parameters with correlation coefficients having absolute values greater than the threshold for at least one listener were selected as relevant for the corresponding scale. The number N_p of parameters selected as relevant by this procedure was 9 for G, 7 for R, 16 for B, 13 for A, and 7 for S.

Tab. II summarises the signs of the relevant correlation coefficients identified by the afore-mentioned procedure. For the average band energy, a negative correlation implies higher rates for lower energies and a positive correlation means higher rates for higher energies. As for decorrelation time, all relevant correlations are negative, which means higher rates for shorter decorrelation times (more instability in energy).

C. PLS modelling

Fig. 1 shows how the fraction of variance in GRBAS rates explained by the PLS model in (1) evolves as the number of latent variables N_f varies. The PLS model has been built using the acoustic features in Tab. II as inputs.

Due to the limited number of input variables, models with more than 7 latent variables did not converge. The graphs in Fig. 1 indicate that a quasi log-linear dependence of the fraction of explained variance from the number of latent variables happens for up to 4 latent variables. Beyond that number, the fraction of explained variance in G, R and S only experiences minor changes and although it has a more relevant growth for B and A, this is still lower than 10% of its value for $N_f = 4$.

Considering the previous observations, a PLS model with 4 latent variables has been selected. For all GRBAS scales, the variable which corresponds to the greatest fraction of variance explained by the model is an average of all relevant acoustic parameters with similar weights for all of them and with weight signs as indicated in Tab. II. For the scales with the highest fraction of variance explained by the model (A and B, as shown in Fig. 1) the weights assigned to the first two variables for the five listeners are similar, while the most relevant differences happen for the third and fourth variables. In contrast, for R and S similarity only happens in the first variable.

Tab. III shows the correlation coefficients between rates assigned by listeners and rates predicted by the PLS models. Not surprisingly, the highest correlations occur for the scales with the highest fractions of variance explained by the model. The mean values of such correlation coefficients, averaged for all scales, are similar for all listeners. These mean values are around 0.6, below the average values of inter-rater correlations in Tab. I.

V. CONCLUSIONS

The spectral analysis reported in [12] indicated that the presence of dysphonia was closely related to low energy in frequencies from 1080 to 2700 Hz and high energy in bands over 5300 Hz for running text recordings. For the same type of recordings, there also was a looser relationship between dysphonia and shorter energy decorrelation times in frequencies from 630 to 2700 Hz.

Results reported here confirm these relations and they indicate that, among the five dimensions included in the GRBAS scheme, the spectral distribution of energy is more closely related with B and A than with the rest. This may be a cue that B and A mainly depend on low-level auditory features such as the ones used here, while rating of G, R and S requires more complex processing.

The characteristics of the linear model built to relate GRBAS rates and acoustic features also revealed that the greatest part of the rate variations explained by the model can be attributed to a few factors common for all listeners.

ACKNOWLEDGEMENTS

This work has been partially financed by the Spanish Government, through project grant number TEC2012-38630-C04-01. Voice recording was carried out in the context of project AIB2010DE-00304, jointly financed by the Spanish Government and the Deutscher Akademischer Austauschdienst (DAAD).

REFERENCES

[1] P.H. Dejonckere, C. Obbens, G.M. DeMoor and G.H. Wieneke, "Perceptual evaluation of dysphonia: Reliability and relevance", *Folia Phoniatr Logop*, vol.45, n.2, pp.76–83, 1993.

[2] G.B. Kempster, B.R. Gerratt, K. Verdolini, J. Barkmeier-Kraemer and R.E. Hillman, "Consensus auditory-perceptual evaluation of voice: Development of a standardized clinical protocol", *Am J Speech-Lang Pat*, vol.18, n.2, pp.124–132, 2009.

[3] R. Buekers, "Perceptual evaluation of vocal behaviour", *Logoped Phoniatr Vocol*, vol.23 (PEVOC-II Suplem.), pp.23–27, 1998.

[4] P. Carding, E. Carlson, R. Epstein, L. Mathieson, C. Shewell, "Formal perceptual evaluation of voice quality in the United Kingdom", *Logoped Phoniatr Vocol*, vol. 25, n.3, pp.133–138, 2000.

[5] M.S. De Bodt, F.L.Wuyts, P.H. Van de Heyning, C. Croux, "Test-retest study of the GRBAS scale: Influence of experience and profesional background on perceptual rating of voice quality", *J Voice*, vol.11, n.1, pp.74–80, 1997.

[6] J. Kreiman, B.R. Gerratt, M. Ito, "When and why listeners disagree in voice quality assessment tasks", *J Acoust Soc Am*, vol.122, n.4, pp.2354–2364, 2007.

[7] T. Bhuta, L. Patrick, J.D. Garnett, "Perceptual evaluation of voice quality and its correlation with acoustic measurements", *J Voice*, vol.18, n.3, pp. 299–304, 2004.

[8] A. McAllister, J. Sundberg, S.R. Hibi, "Acoustic measurements and perceptual evaluation of hoarseness in children's voices", *Logoped Phoniatr Vocol*, vol.23, n.1, pp.27–38, 1998.

[9] R. Shrivastav, C.M. Sapienza, "Objective measures of breathy voice quality obtained using an auditory model", *J Acoust Soc Am*, vol.114, n.4, pp.2217–2224, 2003.

[10] N. Sáenz-Lechón, R. Fraile, J.I. Godino-Llorente, R. Fernández-Baíllo, V. Osma-Ruiz, J.M. Gutiérrez-Arriola, J.D. Arias-Londoño, "Towards objective evaluation of perceived roughness and breathiness: An approach based on mel-frequency cepstral analysis", *Logoped Phoniatr Vocol*, vol.36, n.2, pp.52–59, 2011.

Table III. Values of the Spearman coefficients of correlations between rates assigned by each listener and those predicted by its corresponding PLS model.

	R1	**R2**	**R3**	**R4**	**R5**	**Aver.**
G	*0.62*	*0.57*	*0.48*	*0.63*	*0.55*	0.57
R	*0.57*	*0.52*	*0.45*	*0.63*	*0.50*	0.53
B	*0.64*	*0.53*	*0.80*	*0.65*	*0.68*	0.66
A	*0.68*	*0.75*	*0.61*	*0.71*	*0.67*	0.68
S	*0.69*	*0.56*	*0.53*	*0.57*	*0.41*	0.55
Aver.	0.64	0.59	0.57	0.64	0.56	0.60

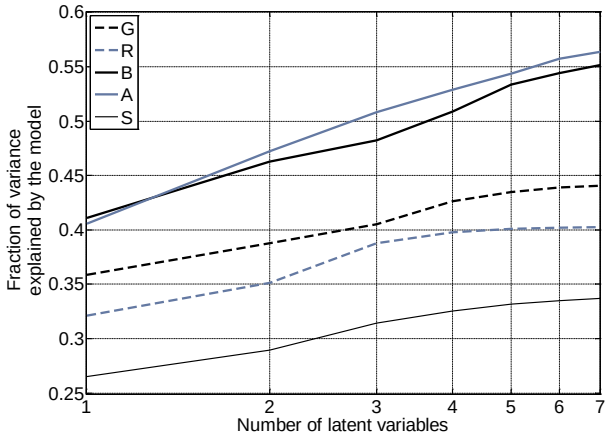

Figure 1. Fraction of variance in GRBAS rates explained by the PLS model vs the number of latent variables N_f.

[11] Y.D. Heman-Ackah, D.D. Michael, G.S. Goding, "The relationship between cepstral peak prominence and selected parameters of dysphonia", *J Voice*, vol.16, n.1, pp.20–27, 2002.

[12] R. Fraile, J.I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruiz, J.M. Gutiérrez-Arriola, "Characterization of dysphonic voices by means of a filterbank-based spectral analysis: Sustained vowels and running speech", *J Voice*, vol.27, n.1, pp.11–23, 2013.

[13] D.M. Howard, E. Abberton, A. Fourcin, "Disordered voice measurement and auditory analysis", *Speech Commun*, vol.54, n.5, pp.611–621, 2012.

[14] H. Hermansky, "Speech representations based on spectral dynamics", *MAVEBA 2013*, pp.191–194, 2013.

[15] S. Orlandi, P.H. Dejonckere, J. Schoentgen, J. Lebacq, N. Rruqja, C. Manfredi, "Effective pre-processing of long term noisy audio recordings: An aid to clinical monitoring", *Biomed Signal Process Control*, vol.8, n.6, pp.799–810, 2013.

[16] M. Slaney, "An efficient implementation of the Patterson-Holdsworth auditory filter bank", *Apple Computer, Perception Group*, Tech rep, 1993.

[17] J.J. Higgins, *An Introduction to Modern Nonparametric Statistics*, Brooks/Cole, 2004.

[18] S. Wold, M. Sjöström, L. Eriksson, "PLS-regression: A basic tool of chemometrics", *Chemometr Intell Lab Syst*, vol.58, n.2, pp.109–130, 2001.